文章编号: 1003-0077(2008)03-0071-10

文本意见挖掘综述

姚天 昉[1],程希文[2],徐飞玉[2],汉思·乌思克尔特[2,3],王 睿[3]

(1. 上海交通大学 计算机科学与工程系,上海 200240;2. 德国人工智能研究中心,德国 萨尔布吕肯 D-66123;3. 德国萨尔州大学 计算语言学系,德国 萨尔布吕肯 D-66041)

摘 要: 意见挖掘是针对主观性文本自动获取有用的意见信息和知识,它是一个新颖而且十分重要的研究课题。这种技术可以应用于现实生活中的许多方面,如电子商务、商业智能、信息监控、民意调查、电子学习、报刊编辑、企业管理等。本文首先对意见挖掘进行了定义,然后阐述了意见挖掘研究的目的,接着从主题的识别、意见持有者的识别、陈述的选择和情感的分析四个方面对意见挖掘的研究现状进行了综述,并介绍了几个成型的系统。此外,我们针对汉语的意见挖掘做了特别的分析。最后对整个领域的研究进行了总结。

关键词: 计算机应用; 中文信息处理; 意见挖掘;主观性文本;综述

中图分类号: TP391 **文献表示码**: A

A Survey of Opinion Mining for Texts

YAO Tian-fang[1], CHENG Xi-wen[2], XU Fei-yu[2], Hans USZKOREIT[2,3], WANG Rui[3]

(1. Dept. of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;
2. German Research Center for Artificial Intelligence, Saarbr cken D-66123, Germany;
3. Dept. of Computational Linguistics, Saarland University, Saarbr cken D-66041, Germany)

Abstract: Opinion Mining is a novel and important research topic, aiming to automatically acquire useful opinioned information and knowledge in subjective texts. This technique has wide and many real-world applications, such as e-commerce, business-intelligence, information monitoring, public-opinion poll, e-learning, newspaper and publication compilation, business management, etc. In this paper, we give a definition for opinion mining and then describe the motivation of this research. Afterwards, we present a survey on the state-of-the-art of opinion mining on top of four subtasks: topic extraction, holder identification, claim extraction and sentiment analysis, followed by an overview of several existing systems. In addition, specific analysis on Chinese Opinion Mining is performed. Finally, we provide the summarization of opinion mining research.

Key words: computer application; Chinese information processing; opinion mining; subjective texts; survey

1 意见挖掘的定义

近年来,对描述非事实(Non-factual)的主观性文本(Subjective Texts)处理方面的研究十分活跃,主要的特点是对基于断言(Allegations)或评论(Comments)的文本进行处理,此类文本内容包含有个人、群体、组织等的意见(Opinions)、情感(Feelings)和态度(Attitudes)等。其中对意见型的主观性文本进行研究形成了一个新颖而且十分重要的研究课题—意见挖掘(Opinion Mining)。

主观性文本是相对于客观性文本而言的一种自然语言文本表达形式。它主要描述了作者对事物、人物、事件等的个人(或群体、组织等)想法或看法。

收稿日期: 2007-06-11 **定稿日期**: 2007-09-29

基金项目: 国家自然科学基金资助项目(60773087)

作者简介: 姚天 昉(1957—),男,博士,副教授,主要研究方向为意见挖掘、信息抽取、机器学习等;程希文(1980—),女,硕士,助理研究员,主要研究方向为意见挖掘,信息抽取等;徐飞玉(1969—),女,博士,高级软件工程师,项目经理,主要研究方向为意见挖掘、信息抽取、机器学习等。

这类文本通常出现在报刊(如读者意见)、互联网(如论坛、电子公告)等媒体上。其中,在文本中包含有表达意见的语句,即具有褒贬意义成分的语句,我们称此类文本为意见型主观性文本。

根据 Kim 和 Hovy 对意见的定义[1]:意见由四个元素组成:即主题(Topic)、持有者(Holder)、陈述(Claim)、情感(Sentiment)。这四个元素之间存在着内在的联系,即意见的持有者针对某主题发表了具有情感的意见陈述。需要指出的是,有时主题也被称为焦点或对象(Focus),以区别可能产生的歧义。

例 1:"我昨天买了一台笔记本电脑,它不但外观漂亮,而且功能强大。"

在上例的句子中,一共有三句子句。第一句是客观句,所以它不是包含意见的陈述。第二、三句子句则是主观句,而且是包含意见的陈述。对于这两个陈述,它们的主题分别是"外观"和"功能",它们都是"笔记本电脑"的性质。这两个陈述的意见持有者是作者"我"。在陈述中所描述的情感是"漂亮"和"强大",都是褒义的。

意见挖掘的过程就是要在自然语言文本中自动确定这些元素以及它们之间的关系。一般来说,它的主要子任务有:

a) 主题抽取(Topic Extraction):识别主题术语和指派领域相关的本体概念;

b) 意见持有者识别(Holder Identification):确定意见表述的作者和谈话者;

c) 陈述的选择(Claim Selection):确定意见表述的范围和过滤客观性表述;

d) 情感分析(Sentiment Analysis):决定意见陈述的语义倾向(Semantic Orientation)(即极性(Polarity))。

意见挖掘是一个多学科综合的研究领域,与文本挖掘、信息抽取、信息检索、机器学习、自然语言处理、概率论、统计数据分析、本体学(Ontology)、语料库语言学、可视化技术等均相关。

2 意见挖掘研究的目的

目前,互联网上的信息与日剧增,蕴藏着巨大的信息量。但是,要想在很短的时间内获得人们对于诸如人物、事件、传媒、产品等有价值的评价信息,往往是十分困难的。例如,对产品的各种评价出现在各大论坛、电子公告板以及门户网站上,厂商需要了解顾客使用其产品的反馈意见,潜在的购买者也需要作出是否购买某个产品的决定。如果采用人工方式对这浩如烟海的信息进行查询、统计,显然是低效和不切合实际的。

面对这样的现实问题,意见挖掘技术应运而生。一方面,它基于数据挖掘(Data Mining)和文本挖掘(Text Mining)技术,另一方面,它又具有相当的文本理解(Text Understanding)的能力。所以,它是比文本挖掘技术更接近人工智能目标的一种新技术。它与以往的信息抽取(Information Extraction)、文本分类(Text Classification)和文本摘要(Text Summarization)技术不同。虽然信息抽取和意见挖掘都需要深层的语义理解,但信息抽取主要是获取具体的语言表达结构,如命名实体、命名实体关系、事件等,这些成分一般为显式表达结构;而意见挖掘是挖掘意见的元素和它们之间的关系,即主题、意见持有者、陈述、情感和它们之间的关系,这些成分表达形式多样,而且常常不是显式地、独立地表达。文本分类是在预定的用户需求下把文本进行分类,并没有涉及到深层次的语义理解。文本摘要是用简练的语言表达长篇文本的中心思想,但文本中对事物的具体看法和评价则没有被清晰地提取出来。实际上,意见挖掘技术弥补了上述这些技术的不足,是更具有应用价值的一种新技术。

意见挖掘涉及各个语言分析层面,不但涉及到词汇层(如分词和词性标注)、句法层(如命名实体识别和语法分析)和语义层(如语义分析),还涉及到篇章层(如跨句的指代消解)。

意见挖掘与一些语言技术有关,例如,信息检索、文本分类、信息抽取、自动摘要、数据融合、问答系统、自然语言生成、对话系统、机器翻译等。

意见挖掘技术可以应用于现实生活中的许多方面,如电子商务、商业智能、信息监控、民意调查、电子学习、报刊编辑、企业管理等。例如,采用意见挖掘系统从来自网上的产品(如笔记本电脑)评价意见中快速地获得意见分类统计结果,可以提供给厂商以进一步改进产品的质量,可以提供给潜在的顾客作为选择购买什么型号产品的参考,也可以提供给代销商作为进货品种和数量的依据。

图 1 说明了在意见挖掘处理中所涉及的语言分析层面、相关语言技术和部分应用领域。

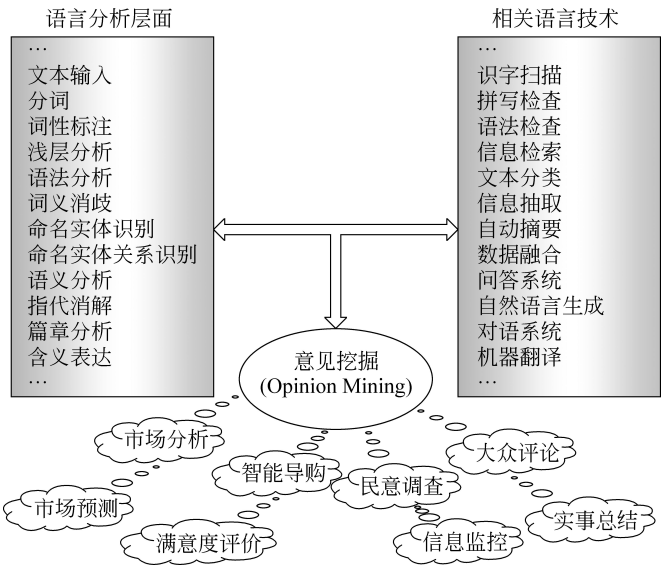

图 1　意见挖掘所涉语言分析层面、相关语言技术和部分应用领域

虽然国际国内的一些研究者已经开展了对意见挖掘技术的研究,也产生了一些应用系统(见下一节"意见挖掘研究现状")。但它还是面临一些尚未解决的问题,如意见挖掘方法的精度和鲁棒性不理想、隐式主题(陈述中不包含具体表示主题的词汇或短语)的识别以及细颗粒度的主题和情感关系识别(如多主题和多情感的对应关系)等。

例 2:"后备箱没有内开装置,我觉得很不安全,也很不爽!"

在例 2 的句子中第一个子句是客观句,第二、三子句是包含意见的陈述。但是这两个陈述中缺乏显式主题(陈述中包含具体表示主题的词汇或短语)。此外,它们的主题不是"内开装置",而是"后备箱"。它对应了两个情感描述项,一个是"很不安全",另一个是"很不爽"。注意这两个情感描述项中包含有否定词。所以,它们的语义倾向与"安全"和"爽"相反,是贬义的。再加上修饰词"很",加强了贬义语义倾向的强度。另外,从第三句子句中的"爽"字,可以知道情感描述项不但涉及书面表达形式,也涉及口语表达形式。

综上所述,意见挖掘是一种很有前途的新颖的语言技术。但是,它同时也面临着许多尚未解决的问题和挑战。因此,对于意见挖掘方法和技术的研究,不但具有理论意义,而且具有实用价值。

3　意见挖掘研究的现状

在这节中,我们首先从主题的识别、持有者的识别、陈述的选择和情感的分析这几个方面介绍国际上对于意见挖掘的相关方法和技术的研究现状,然后介绍几个成熟的意见挖掘应用系统。最后介绍国内意见挖掘方法和技术的研究现状,主要是针对汉语的特点探讨意见挖掘的方法和技术。

3.1　主题的识别

为了清楚地了解识别主题的任务,请参考如下例子:

例 3:"诺基亚的机器品质一直不错,而且 6101 也是一款经过市场考验的机器,现在的价格可以说是非常合理的。"

第一个子句的主题是"品质",第二个子句的主题是"6101"(即诺基亚 6101 型手机),但第三个子句的主题是"价格"。这些主题都是显式主题。因为大多数研究都是面向产品的意见挖掘,显式主题主要是领域相关的术语。在这方面,主要有两种策略:第一是根据短语结构的特点。Yi 等人根据名词短语的组成和位置特点,采用相似性测试(Likelihood test)方法来确定主题[2]。但是,这种方法不能有效的解决主题术语的覆盖性问题;第二是根据候选主题的同现(Co-occurrence)和上下文指示符。Hu 和 Liu 根据主题和一些指示词的共显特征来识别常现(Frequent)和非常现(Infrequent)主题术语[3]。与此相似的是 Popescu 和 Etzioni 采用点互信息(Point-wise Mutual Information, PMI)方法获取候选主题术语[4]。候选主题术语的 PMI 值是它与产品的整体—部分关系识别符(Meronymy Discriminator),比如对于扫描仪的识别符是 scan of, scan'

s,等)组之间的互信息(MI)值。这个互信息不单单可以在有限训练数据中获得,也可以从整个互联网获得。实验证实:第一种方法虽然有较高的查全率,但是第二种方法的查准率远远高于第一种方法。

并非所有的句子主题都是显式的,在例2中我们也看到了隐式主题。显然,对隐式主题的识别要依赖上下文的语义分析,具有相当的难度。目前大多数主题识别的流行方法主要涉及显式主题的识别,因此这将成为我们的研究重点之一。另外,我们可以看到例子中"品质"、"6101"和"价格"分别代表手机品质、手机型号和手机价格。在大多数情况下,这些作为主题的领域术语主要针对的是所评价产品的成分、性质和功能。为了能够对所评价的对象进行全面地、正确地评价,精确地给所评价主题指派(Assignment)特定领域的本体概念是非常重要的,例如,"品质"和"质量"就是相同的概念。这些针对概念的意见总结能提供给关心者最全面的信息。

本体概念指派是一个新的研究领域,在这方面还只有很少的研究工作。文献[4]利用WordNet的IS-A层次关系和词形态线索(例如,"-iness"、"-ity"后缀)来分类主题术语的"part-of"和"property-of"类别。但是因为WordNet的局限性,这种解决方案不能有效地识别领域相关词汇,比如缩写词、专业行话(Jargon)等。针对领域词汇的特点,Cheng研究了基于本体的主题抽取[5]。为了达到高覆盖率,她首先采用半自动的方法基于一些现存的本体资源构造应用领域本体(汽车领域本体),然后将基于规则的命名实体识别技术和信息抽取引擎结合起来,识别被评价的汽车领域主题术语并给它们指派该应用领域的相关概念。

3.2 意见持有者的识别

这一节的介绍从分析下面这个例子开始。

例4:"科学家们相信这项研究会产生深远的影响。"

在上例中,"科学家们"是关于这项研究的意见持有者。直观上看,识别意见持有者的研究可采用命名实体方法[1],即考虑把人名或机构名作为可能的意见持有者。此外,语言资源也能被用来解决这个问题。例如,FrameNet可以被用来抽取作为意见持有者的特定语义角色[6,7]。然而,利用自然语言处理(Natural Language Processing, NLP)技术作为解决方案会产生两种可能的缺点:较低的语言现象覆盖率和较差的领域适应性。另外,这个意见挖掘的子任务也能通过基于模式识别的信息抽取(Information Extraction, IE)和机器学习(Machine Learning, ML)技术来解决[8]。但是,到目前为止,这些技术还不能解决识别意见持有者之间的关系。请看下面的例子:

例5: Sergey Stanishev表示:"保加利亚加入欧盟的推迟将是一个严重的错误"。自从Stanishev在2005年被选举担任保加利亚总理以来,他一直是他的国家加入欧盟的著名支持者。

在上述两句语句中,意见持有者是Sergey Stanishev,即保加利亚总理。他针对"保加利亚加入欧盟的推迟"持反对意见(贬义),而对"他的国家加入欧盟"持赞成意见(褒义)。在这种情况下,如果我们要知道意见持有者和意见之间准确的联系,重要的是要解决意见持有者所涉及的互指(Co-reference)问题。实际上,文献[9]把意见持有者与真实世界的实体关联起来,根据选择提到相同实体的意见持有者来解决互指问题。实验结果与当前名词短语互指消歧的结果相当。识别意见持有者在新闻文本的观点总结中是特别重要的。但是在产品评价文本中,它们的意见持有者可被考虑为一个评价人组。

3.3 陈述的选择

陈述是一种具有情感的意见表述。所以,它是属于主观性的表述。如果不区分主客观语句的话,将会给情感分析带来干扰。下面就怎样识别陈述以及怎样进行主观性分析介绍一下有关的研究工作。

例6:引用专家评论:"长期以来帕萨特是大众汽车中最好看的车型之一,它的内部非常精致。它在美国交易艰难的中型私家车市场上一直是持续最畅销的德国汽车。这款汽车装备了新的悬挂。停车距离合适,但刹车踏板太软。像以往一样,新款式将吸引许多认为德国制造了最好汽车的美国人。"

针对不同颗粒度意见挖掘的需要,陈述颗粒度也有不同的大小。一般来说,流行的意见挖掘方法中,陈述的颗粒度可分成粗颗粒度、细颗粒度和特定颗粒度。而且客观性语句"这款汽车装备了新的悬挂"必须被过滤,因为它的存在干扰了意见挖掘的处理。通常,主观性分析可转换为两元分类的文本分类任务。陈述范围的颗粒度越是精确,完成这个任务的难度也越大。

第一,大多数研究工作进行了粗颗粒度陈述的研究,即假定一个文档或一句语句是一个关于给定

或被识别主题的陈述。例如,在例 6 中,关于主题“2007 年大众帕萨特”的最粗颗粒度的陈述包含了整个文档。另一种比较特定的解决方案是分析每一句语句的情感以及所有共享相同语义倾向(极性)的被包含的主题,以这样的语句作为一个陈述。举例来说,语句“长期以来帕萨特是大众汽车中最好看的车型之一,它的内部非常精致。”的主题“帕萨特”和“内部”共享相同的情感极性“褒义”。那么,这就是一个陈述。但是,这种情况不总是正确的,如其后的一句语句“停车距离合适,但刹车踏板太软。”主题“停车距离”和“刹车踏板”具有不同的情感极性,不符合这个解决方案,也就不能成为一个陈述。显然,定义整个文档或一句语句作为一个陈述的范围太一般性了,以至于不能获得正确的意见。

对于粗颗粒度的主观性分析,文档或语句被编码为词袋(Bag-of-words),一些特征被证明是有用的: 如词典计数[10]、主观性词汇和短语[11,12]、语义倾向(Semantic Orientation, SO)[13]、语义组块程序[12]或它们的组合[14,15]。分类器的相关的算法包括 K 近邻(K-nearest)[11]、贝叶斯[15]、筛选(Winnow)[12,14]等。此外,除了经典的机器学习算法,Pang 和 Lee 采用基于寻找文档的最小图割(Minimum Cuts in Graphs)的方法来寻找文档中的主观性部分[10],即识别文档中语句与已知主观语句的对应关系,从而达到对主客观语句分类的目的。这个方法能够被成功地应用于语句层和文档层的主观性分析。

第二,细颗粒度陈述着眼于更特定的范围,从而可以抽取主题和陈述之间紧密的关系。如定义包含形容词或副词的名词短语作为陈述[16]。在例 6 中,短语例如“持续最畅销的德国汽车”和子句例如“但刹车踏板太软”、“它的内部非常精致”是表达意见的细颗粒度陈述。Wilson 等人开创了在短语层进行主观性分析的研究工作[17]。他们根据 28 个混合特征训练了一个 BoosTexter 分类器。这些混合特征被分成五个类别: 词汇特征(例如,上下文、原始极性)、修饰特征(形容词开始、修饰强的主观性指示符)、句子特征(例如,当前句中强的主观性指示符)、结构特征(例如,被动态)和文档特征(主题)。实验结果: 这个五特征集合的分类器在细颗粒度主观性分析任务中完成得很好。

第三,还有一些方法[2,12]使用特定手工编辑的语言模式来从评论中抽取陈述。例如,对于语句“新款式将吸引许多认为德国制造了最好汽车的美国人。”,采用语言模式〈吸引,主语(主题+),宾语(意见持有者)〉来抽取陈述。这个语言模式表示作为宾语的意见持有者对于作为主语的主题有一个褒义的意见。然后,被抽取的陈述就是〈吸引,(新款式),(美国人)〉,它指出了美国人针对主题“新款式”有一个褒义的意见。这种陈述的范围是非常严格的并达到了较高的精度。但是,它难以改进模式的覆盖面和可适应性。

3.4　情感的分析

心理学研究[18]发现,词汇和人类情感之间的关系是可度量的,独立的词汇或短语的语义倾向对于传达人类情感是重要的。请看以下的例子:

例 7:“气候宜人的旅游胜地和引人入胜的山山水水。”

在上例中,形容词“气候宜人”和“引人入胜”分别表达了对主题“旅游胜地”和“山山水水”褒义的情感。直观地看,对每一个陈述,总的语义倾向好/坏量值被用来表示意见持有者的情感。现今的研究重点是语义倾向的自动获取。针对语言的特色,有两个现象被广泛地用来获取词汇和短语的语义倾向: 第一,相同倾向的情感术语经常同时出现[16,19];第二,相反倾向的情感术语一般不一起出现[20]。根据这两个现象,可以仅使用很少的种子(例如,“excellent”是褒义的,“poor”是贬义的)从一个大型语料库中(例如,互联网)获得词汇和短语的情感倾向。为了发现更多的情感词汇和短语,获得情感倾向的词和短语能够被加入到种子集中进行迭代寻找[15,20]。除了语言的特点,已存在的资源所起的作用也十分显著。比如,WordNet 中词汇之间的距离可以揭示情感倾向的关系[21]。此外,其中形容词的同义词和反义词集也可以被用来预测情感倾向[1,10]。

然而,情感被传递时不仅通过单一词汇或短语,而且也可能通过它们的组合。根据下面的例子,即使形容词“满意”传递褒义的极性,但是整个陈述表达了相反的极性。

例 8:“我们对于售后服务不是很满意。”

这个陈述的情感由词汇的组合“不是很满意”所表达。词汇“不是”传递了词汇“满意”的相反极性。这样,像“不是”这样的词汇属于配价移动指示符(Valence Shifter Indicator,VSI)。即使配价移动指示符不具有语义倾向,它们也能传递对应情感倾向总的量值的相反极性。因此,我们应该考虑超出语义倾向更多的限制。目前,进一步的研究工作不仅专注于个别词汇和短语,也需要专注于陈述中多个

元素之间的相互作用。比如,一些方法采用了手工构造语言模式或赋予陈述极性的限制[2,12]。例如,在例8中,语言模式〈对于 POBJ (-)不是很满意〉能被容易地用来区别已填入槽〈POBJ (-)〉中的主题"售后服务"是贬义的。这个方法产生了相当高的精度。但是,它要求许多手工的工作而且语言现象的覆盖面不广。

另一个方法专注于上下文各元素之间的相互作用[4]。这个过程首先获取词汇的语义倾向,然后根据手工编辑的一些限制(例如,并列关系连接词连接相同语义倾向的词)给元组〈word〉、〈word,topic〉和〈word,topic,sentence〉以迭代方式指派极性。例如,在例8中,"满意"通过句法依赖规则被选择为情感词汇,然后通过 PMI 算法,情感词汇"满意"开始时被指派一个极性。根据限制条件,并列关系连接词"和"联系具有相同极性的两个形容词,所有的同义词具有相同的极性。这样,产生了第一个元组(a)〈满意〉,它具有正的极性,接着,根据这个算法,第二个元组(b)〈#a,售后服务〉的极性等于第一个元组的极性,它也能够根据上下文指示符的限制被改变(例如,并列或转折关系连接词)。相似的策略也被应用到第三个元组(c)〈#a,#b,我们对于售后服务不是很满意〉。由于词汇"不是"是 VSI,第三个元组的极性与第二个元组相反。这个迭代过程直到在语料库中元组的极性不能被改变为止。这种方法在大量手工限制干预下,能够达到较高的准确率。

上述的语言模式和限制都需要大量人工信息的干预,因此为了提高技术的领域实用性,另外的研究把这些信息都作为特征去训练情感统计模型。在这方面,我们主要介绍对于不同陈述范围所选取的特征。文档层模型集成了 N 元特征,达到了较高的性能[22,23]。但是,这些模型在句子层分析效果较差[23],它们需要合并针对句子层分类的附加特征。除了 N 元模型之外,从组成或依赖结构抽取的子串改善了句子层模型的性能[24,25]。由于在句子中的稀疏信息,他们认为集成更多的特征是必要的[26]。

最近,一些研究工作已经开始探索更细颗粒度的情感分析[17]。他们使用、收集和混合所有提到的特征(例如修饰语的极性、通常的 VSI 词汇像"little threat"、否定极性移动符像"lack of understanding")来指派极性给短语和子句。实验证实了许多语言信息的合并显著地改进了细颗粒度的情感分析的性能。Cheng 研究了细颗粒度的情感分析[5]。她采用机器学习方法从单一领域语料库中学习获取独立词汇和 VSI 的语义倾向。根据主题的本体概念和语义倾向使用启发式规则选择陈述。在此基础上,使用统一的表示集成具有情感知识的语言特征,然后采用朴素贝叶斯分类器分类情感极性。

3.5 意见挖掘应用系统

意见挖掘技术作为一种新颖的语言技术不仅可以运用于自然语言处理系统中,还可以应用于现实生活中。例如:

Dave 等人研究并开发的 ReviewSeer 是世界上第一个情感分析工具和第一个针对给定产品评论区别其褒贬性的系统[23]。Gamon 等人研究并开发的 Pulse 系统可以自动挖掘网上用户所上载的自由文本中有关汽车评价中的贬褒信息和强弱程度[27]。Liu 等人研究并开发的 Opinion Observer 系统可以处理网上在线顾客产品评价[28],对涉及产品(如电子照相机)各种特征的优缺点进行统计,并采用可视化方式对若干种产品特征的综合质量进行比较。Yi 和 Niblack 研究并开发了 WebFountain 系统中的意见挖掘器[29]。WebFountain 系统是一个基于多类型数据、开放领域意见挖掘的研究和开发平台。Wilson 等人研究并开发的 OpinionFinder 是一个自动识别主观性句子以及句子中各种与主观性有关的成分(例如,意见源、直接的主观性表达、说话事件(Speech Event)、情感等)的系统[30]。

最后,为了更清楚地比较各个应用系统所挖掘的意见元素、情感分析类型以及情感知识使用情况,列表如下:

表1 各个应用系统所挖掘的意见元素、情感分析类型以及情感知识使用情况的比较

应用系统名称	所挖掘的意见元素	情感分析类型	情感知识使用情况
ReviewSeer	〈陈述,情感〉	粗颗粒	未使用
Pulse	〈陈述,主题术语,情感〉	粗颗粒	使用
Opinion Observer	〈陈述,主题术语,情感〉	粗颗粒	使用
WebFountain	〈意见持有者,陈述,主题术语,情感〉	特定颗粒	使用
OpinionFinder	〈意见持有者,陈述,主题术语,情感〉	细颗粒	使用

3.6　汉语意见挖掘研究现状

汉语意见挖掘方法和技术的研究起步较晚。

Yuan 等人在 Turney 和 Littman 研究工作[31]的基础上,对汉语极性词的自动获取进行了研究[32]。研究发现,采用一个字符的汉语极性词素(Morpheme)比用汉语极性词作为极性基准词素(词)(比较对象)的效果要好。与 Turney 的研究结果相比,汉语语料库的规模大大小于 Turney 和 Littman 所需要的语料库规模。

此后,Tsou 等人在上述研究工作基础上对中国四地(北京、香港、上海、台北)报刊上有关四位政治人物(克里、布什、小泉纯一郎、陈水扁)褒贬性的新闻报道进行了分类研究[33]。在研究中,首先通过标记语料库获得文本中的极性元素(Polar Elements),然后主要采用了三个衡量指标,即极性元素的散布(Spread)、极性元素的密度(Density)和极性元素的语义强度(Intensity)来对每个文本进行统计,得出文本贬褒分类和强度大小的结果。其中对确定极性元素之间的关系虽有所提及,但没有深入研究。

在 BBS 文本研究方面,邱立坤等人提出了一种在 BBS 环境下进行热门话题挖掘的算法[34]。这种算法在一般文本聚类算法基础上,应用 BBS 所特有的点击数、回复数进行热度排序,然后采用基于特征词提取的话题归并,从而挖掘出最受 BBS 用户关注的热门话题。

在主观性语言研究方面,Xia 等人对于汉语网络非正规语言(Network Informal Language,NIL)(例如,NIL 中的"偶"等同于"我")进行了研究[35]。他们利用 BBS 文本建立了 NIL 语料库,通过对语料库中语料进行观察。他们认为该语料具有奇异性(相对于标准词汇的奇异效果,例如,"稀饭"代表"喜欢")和动态性(NIL 变化很快,无法用静态词典去覆盖)的特点。为了解决动态性问题,他们采用了语音映射模型(Phonetic Mapping Model)去完成 NIL 词汇到标准词汇的映射,即通过拼音实现语音转换。为了完成 NIL 的标准化,他们通过合并语音映射模型来扩展源通道模型(Source Channel Model)。实验表明,这种方法对于动态 NIL 的标准化是有效而稳定的。

在汉语词汇语义倾向自动获取研究方面,朱嫣岚等人提出了基于 HowNet 的两种词汇语义倾向计算方法[36],即基于语义相似度的方法和基于语义相关场的方法。实验表明,在同一测试集上,基于 HowNet 语义相似度的方法比基于语义相关场的方法精确率高。

在汉语句子语义极性分析和观点抽取研究方面,娄德成和姚天　利用自然语言处理技术,对汉语网络评论的语句进行了语义极性分析和观点抽取[37]。提出了计算词语的上下文极性的算法,并且分析了主题和极性修饰成分的匹配关系。在观点抽取方面,采用 SBV 算法加上补充算法对主题和极性修饰成分的匹配关系进行了识别。经过与手工标注的结果进行比较,显示了较好的性能。

在汉语文本语义倾向自动识别方面,徐琳宏等人提出了一种基于语义理解的识别机制[38]。首先计算文本中词汇与 HowNet 中已标注褒贬性词汇间的相似度,获取词汇的语义倾向性。在此基础上,选择语义倾向性明显的词汇作为特征值,用 SVM 分类器分析文本的褒贬性。最后采用否定规则匹配文本中的语义否定以提高分类效果,以及处理程度副词附近的褒贬义词以加强对文本褒贬义强度的识别。

在应用系统方面,姚天　等人研究并开发了用于汉语汽车论坛的意见挖掘系统[39]。该系统可在电子公告板、门户网站的各大论坛上挖掘并且概括意见持有者对各种汽车品牌的不同性能指标的评论和意见,并且判断这些意见的褒贬性以及强度。最后,通过对文本处理的综合统计,并给出可视化的结果。

4　结论

上面介绍了国内外在意见挖掘方法、技术和应用研究中所取得的成果。在这些研究中,主要是围绕着四个子任务展开的。其中对主题抽取和情感分析的研究显得尤为重要,因为它们以及它们之间的关系是意见中的核心元素和关系。所以,在这方面的研究工作比较集中。

虽然在国内外所进行的意见挖掘方法、技术和应用研究中已取得了相当的进展,但还存在着一些不足之处,主要是:

(1) 由于意见型主观性文本在书写时随意性较大,遣词造句基本没有约束。对意见挖掘进行研究,首先应从语言学的角度对这种主观性的文本进行研究,包括语料收集、分析语言规律、研究标注规范和方法等。但在国内外的研究中,这方面的基础研究还做得很少。

(2) 目前大部分的意见挖掘方法还是粗颗粒度的。这种方法所造成的问题有：主题术语识别和情感分析等的精度和鲁棒性不如人意；主题术语结构组织不清楚；隐式主题的识别研究欠缺；所涉及的主题和情感之间关系主要是简单关系(如一对一的关系)等。这种粗颗粒度的意见挖掘结果不能直接满足用户细颗粒度的意见挖掘需求，如关于多主题的情感关系以及相应的语义倾向和强度等。

(3) 由于研究的深度有限，目前意见挖掘的一些方法和技术还很难与其他有关的方法和技术相比较，如传统的关系抽取任务之一：事件抽取。

(4) 对汉语文本意见挖掘方法的研究起步于近两年，所做的研究工作还不多。还面临缺乏汉语主观性文本语料库、缺乏语料标注规范、缺乏汉语文本意见挖掘方法的系统性研究、汉语意见挖掘商用系统也未见报道等。

因此，结合我们从事意见挖掘研究工作的经历[5,37,39]，深感上述问题的存在使得意见挖掘方法和技术与实际应用还相距较远。为了能够弥补上述提到的研究中的不足并缩短国内与国外在这方面研究的差距，特别是在构建汉语主观性文本语料库、建立汉语文本意见挖掘计算模型、实现处理大规模汉语真实文本的应用系统等方面加强研究，早出成果，以适应我们国家经济和社会对这方面日益增长的应用需求，无疑是十分重要和必要的。

参考文献：

[1] S.-M. Kim and E. Hovy. Determining the Sentiment of Opinions [A]. In: Proceedings of COLING-04, the Conference on Computational Linguistics (COLING-2004) [C]. Geneva, Switzerland: 2004, 1367-1373.

[2] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques [A]. In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003) [C]. Melbourne, Florida: 2003, 427-434.

[3] M. Hu and B. Liu. Mining Opinion Features in Customer Reviews [A]. In: Proceedings of Nineteeth National Conference on Artificial Intellgience (AAAI-2004) [C]. San Jose, USA: 2004.

[4] A.-M. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews [A]. In: Proceedings of HLT-EMNLP-05, the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing [C]. Vancouver, Canada: 2005, 339-346.

[5] X. Cheng. Automatic Topic Term Detection and Sentiment Classification for Opinion Mining [D]. Master Thesis. Saarbr cken, Germany: The University of Saarland, 2007.

[6] S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky. Automatic Extraction of Opinion Propositions and their Holders [A]. In: J. G. Shanahan et al. (eds). Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications [C]. Stanford, USA: 2004.

[7] S.-M. Kim and E. Hovy. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text [A]. In: Proceedings of the Workshop on Sentiment and Subjectivity in Text (COLING-ACL 2006 Workshop) [C]. Sydney, Australia: 2006, 1-8.

[8] Y. Choi, C. Cardie; E. Riloff and S. Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns [A]. In: Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005) [C]. Vancouver, Canada: 2005, 355-362.

[9] V. Stoyanov and C. Cardie. Toward Opinion Summarization: Linking the Sources [A]. In: Proceddings of the Workshop on Sentiment and Subjectivity in Text (COLING-ACL 2006 Workshop) [C]. Sydney, Australia: 2006, 9-14.

[10] B. Pang and L. Lee. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimun Cuts [A]. In: Proceedings of the ACL-2004 [C]. Barcelona, Spain: 2004, 271-278.

[11] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning Subjective Language [A]. Technical Report TR-02-100 [C]. Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania, USA: 2002.

[12] T. Nasukawa and J. Yi. Sentiment Analysis: Capturing Favorability using Natural Language Processing [A]. In: Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003) [C]. Sanibel Island, Florida, USA: 2003, 70-77.

[13] V. Hatzivassiloglou and J. Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity [A]. In: Proceedings of 18th International Conference on Computational Linguistics (COLING-2000) [C]. New Brunswick, NJ, USA: 2000.

[14] K. Nigam and M. Hurst. Towards a Robust Metric of Opinion [A]. In: Proceedings of the AAAI Spring

Symposium on Exploring Attitude and Affect in Text: Theories and Applications [C]. Standford, USA: 2004.

[15] H. Yu and V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences [A]. In: M. Collins and M. Steedman (eds): Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing [C]. Sapporo, Japan: 2003, 129-136.

[16] P. D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews [A]. In: Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics [C]. , USA: 2002, 417-424.

[17] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis [A]. In: Proceedings of Human Language Technologies Conference/ Conference on Empirical Methods in Natural Language Processing (HLT/ EMNLP 2005) [C]. Vancouver, Canada: 2005, 347-354.

[18] M. M. Bradley, and P.J. Lang. Affective Norms for English Words (ANEW): Stimuli, Instruction Manual and Affective Ratings [A]. Technical report C-1 [C]. Gainesville, FL. The Center for Research in Psychophysiology, University of Florida, Florida, USA: 1999.

[19] P. D. Turney and M. L. Littman. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus [A]. Technical Report ERC-1094 (NRC 44929) [C]. National Research Council of Canada: 2002.

[20] M. Gamon and A. Aue. Automatic Identi? cation of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms [A]. In: Proceedings of the ACL- 2005 Workshop on Feature Engineering for Machine Learning in NLP [C]. Michigan, USA: 2005, 57-64.

[21] J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke. Words with Attitude [A]. In: Proceddings of the 1st International Conference on Global WordNet [C]. Mysore, India: 2002.

[22] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques [A]. In: Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing [C]. Philadelphia, USA: 2002, 79-86.

[23] K. Dave, S. Lawrence, D. M. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classi? cation of Product Reviews [A]. In: Proceedings of the 12th International World Wide Web Conference (WWW2003) [C]. Budapest, Hungary: 2003.

[24] T. Kudo and Y. Matsumoto. A Boosting Algorithm for Classification of Semi-structured Text [A]. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2004) [C]. Barcelona, Spain: 2004, 301-308.

[25] S. Matsumoto, H. Takamura, and M. Okumura. Sentiment Classi? cation using Word Sub-sequences and Dependency Sub-trees [A]. In: Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-05) [C]. Hanoi, Vietnam: 2005, 301-310.

[26] T. Mullen, and N. Collier. Sentiment Analysis using Support Vector Machines with Diverse Information Sources [A]. In: Dekang Lin and Dekai Wu (eds.): Proceedings of EMNLP-04, 9th Conference on Empirical Methods in Natural Language Processing [C]. Barcelon, Spain: 2004, 412-418.

[27] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining Customer Opinions from Free Text [A]. In: Proceedings of IDA-05, the 6th International Symposium on Intelligent Data Analysis [C]. Lecture Notes in Computer Science, Springer-Verlag. Madrid, Spain: 2005, 121-132.

[28] B. Liu, M. Hu, and J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web [A]. In: Proceedings of WWW '05, the 14th international conference on World Wide Web [C]. Chiba, Japan: 2005, 342-351.

[29] J. Yi and W. Niblack. Sentiment Mining in WebFountain [A]. In: Proceedings ICDE-05, the 21st International Conference on Data Engineering [C]. IEEE Computer Society. Tokyo, Japan: 2005, 1073-1083.

[30] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, JanyceWiebe, Yejin Choi, Claire Cardie, Ellen Riloff, Siddharth Patwardhan. OpinionFinder: A System for Subjectivity Analysis [A]. In: Proceedings of HLT/ EMNLP 2005 Demonstration Abstracts [C]. Vancouver, Canada: 2005, 34-35.

[31] P. D. Turney and M. L. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association [J]. ACM Transactions on Information Systems (TOIS), 2003, 21(4): 315-346.

[32] R. Yuan et al. Morpheme-based Derivation of Bipolar Semantic Orientation of Chinese Words [A]. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004) [C]. Ge-

neva , Switzerland: 2004 , 1008-1014.

[33] B. Tsou , R. Yuen , O. Kwong , T. Lai , and W. Wong. Polarity Classification of Celebrity Coverage in the Chinese Press [A]. In: Proceedings of the International Conference on Intelligence Analysis [C]. McLean , USA: 2005.

[34] 邱立坤,程薇,龙志祎,孙娇华. 面向 BBS 的话题挖掘初探 [A]. 见: 孙茂松, 陈群秀主编, 自然语言理解与大规模内容计算 [C]. 北京: 清华大学出版社, 2005 , 401-407.

[35] Y. Xia , K.-F. Wong , and W. Li. A Phonetic-Based Approach to Chinese Chat Text Normalization [A]. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006) [C]. Sydney , Australia: 2006 , 993-1000.

[36] 朱嫣岚,闵锦,周雅倩,黄萱菁,吴立德. 基于 HowNet 的词汇语义倾向计算 [J]. 中文信息学报,2006 , 20 (1): 14-20.

[37] 娄德成 , 姚天 . 汉语句子语义极性分析和观点抽取方法的研究 [J]. 计算机应用, 26 (11) , 2006 , 2622-2625.

[38] 徐琳宏,林鸿飞,杨志豪. 基于语义理解的文本倾向性识别机制 [J]. 中文信息学报, 2007 , 21 (1): 96-100.

[39] 姚天 ,聂青阳,李建超,李林琳,娄德成,陈珂,付宇. 一个用于汉语汽车评论的意见挖掘系统 [A]. 见: 曹右琦,孙茂松主编,中文信息处理前沿进展 —中国中文信息学会二十五周年学术会议论文集 [C]. 北京: 清华大学出版社 , 2006 , 260-281.

(上接第 70 页)

式,建立主题词与网络事件的有机联系,根据相关主题词互信息强的特点,我们设计了新的词聚类策略对其进行组织,使得叙述同一事件的主题词凝聚成簇,组织的效果令人满意。实验表明,我们设计的主题词提取算法是有效的,我们精心设计的组织方式是恰当的。这表明对网络文本主题词的提取与组织不但可行,还能达到良好的实际效果。在下一步的工作中,我们将进一步在滤噪和词串规整方面改进我们的提取算法,在主题词的组织方面,我们将研究簇内更为精细的结构。

参考文献:

[1] Zhang Hua-Ping , Liu Qun , et al. Chinese name entity recognition using role model [J]. Special issue "Word Formation and Chinese Language processing" of the International Journal of Computational Linguistics and Chinese Language Processing , 2003 , 8(2): 29-60.

[2] Wu An-Di , Jiang Zi-Xin. Statistically-enhanced new word identification in a rule-based Chinese system[A]. In: Proc. of the Second Chinese Language Processing Workshop[C]. Hong Kong , China: 2000. 46-51.

[3] Li Hong-Qiao. etc al. The use of SVM for Chinese new word identification[A]. In: Proc. of First International Joint Conference on Natural Language Processing[C]. Sanya , Hainan Island , China: 2004. 497-504.

[4] Ukkonen E. On-line construction of suffix trees[J]. Algorithmica 1995 , 14 , 249-260.

[5] 刘挺,吴岩,等. 串频统计和词形匹配相结合的汉语自动分词系统[J], 中文信息学报,1998 ,12(1): 17-25.

[6] 邹纲. 中文新词自动检测研究[D]. 北京:中科院计算技术研究所 , 2004.

[7] 刘远超,王晓龙,等. 文档聚类综述[J]. 中文信息学报,2006 ,20(3): 55-62.

[8] Dubes R. C. and Jain A. K., Algorithms for Clustering Data[M]. Prentice Hall , 1988.

[9] Kaufman L. and Rousseeuw P. J., Finding Groups in Data: an Introduction to Cluster Analysis[M]. John Wiley and Sons , 1990.

[10] Steinbach M., Karypis G., and Kumar V. A comparison of document clustering techniques [A]. KDD Workshop on Text Mining[C]. 2000.

[11] 周昭涛, 文本聚类分析效果评价及文本表示研究 [D]. 北京:中科院计算技术研究所 , 2005.